1. Utilizaria la métrica de jaccard para poder dividir el vector en conjuntos, que actuarian como shingles (serían grupos de notas, por ejemplo k “coordenadas” del vector, llamo k-subgrupos de notas).
   Las restricciones se dan para los hashes, ya que la longitud de el vector es variable, para resolver esto utilizaria la función de hashing universal con un tamaño fijo (grande) para que no haya problemas con el tamaño del vector de entrada, y para que no haya problemas al acceder a una coordenada del vector de notas que podría no existir, agregaria padding en caso de ser necesario.

   El minhash sería del estilo:
   (a * x[1] + b * x[2] … + z * x[k]) mod p mod m

   Donde m es el tamaño de la tabla, p un número primo mayor a m y a,b,c van de {1, p-1} y son a elección. El k es a definir con el tamaño fijo mencionado arriba

2. Se debe utilizar la familia con parámetros *H(0.2,0.8,0.88,0.16)*.
   El 0.16 es la probabilidad de obtener una nota disímil en una consulta

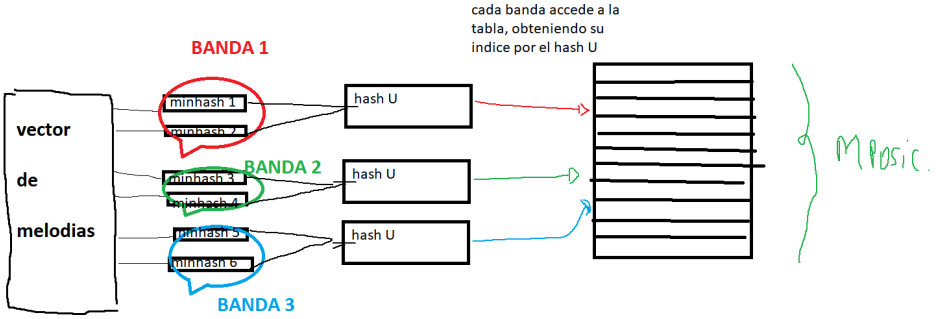
   Donde para amplificarla y cumplir con lo pedido se debe obtener un **b=3** y **r=2**. Con lo cual la cantidad de minhashes va a ser b.r = 6. Es decir, usando 2 tablas (r) y 3 bandas (b) donde por cada banda habrá 2 minhashes. En el ejemplo del punto tres, lo voy a hacer en una sola tabla.

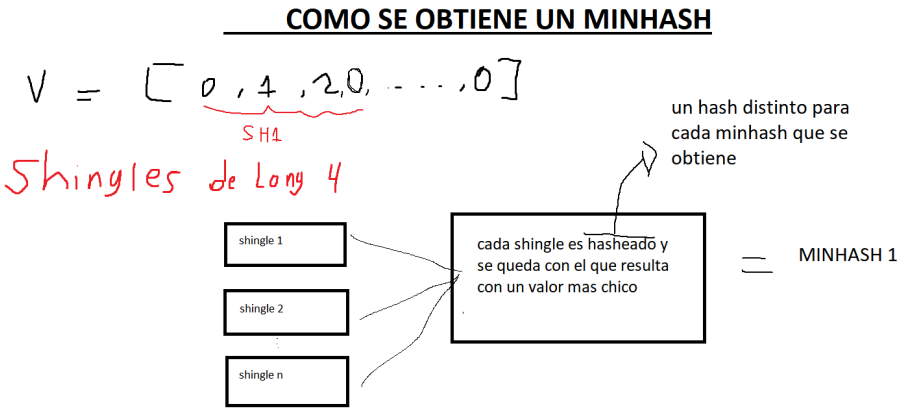
   Con esos valores obtuve p1 = **0.953** y p2 = **0.11**
   Con estas probabilidades se obtiene 0.953 de probabilidad de que objetos similares colisionen y 1-0.953 probabilidad de que objetos similares no colisionen (falsos negativos). Mientras que 0.11 es la probabilidad de que objetos disímiles colisionen.

   El minhash se obtiene de cada banda, calcula un hash para los k-subgrupos que se definen del vector de notas, obteniendo un número y quedándose con aquel hash que obtuvo el valor más chico de la banda.
   Ese resultado, del minhash de cada banda, pasa a otro hash (hash u) el cual devolverá el índice en el cual se guardará la id del vector de notas en la tabla. Este último hash es el siguiente:

   h * mh_1_de_la_banda + f * mh_2_de_la_banda mod p mod m

3.





En la última imagen tome un valor de long de shingles arbitrario. Algunas cosas de la imagen las describí en el punto 2.

4. Para obtener los posibles vectores similares a un vector de consulta, lo que se hace es, aplicar el mismo procedimiento que se muestra en la imagen a ese vector, con lo cual nos dará 2 accesos (b) a la tabla, donde en los buckets de esos índices pueden llegar a encontrarse las id's de las notas similares. Podríamos calcular la semejanza de la nota query y las que obtenemos de los dos buckets, quedándonos con aquella que sea más alta.